УДК 519.6

ИССЛЕДОВАНИЕ ВЛИЯНИЯ СПОСОБА ФОРМИРОВАНИЯ ОБУЧАЮЩИХ ПОДМНОЖЕСТВ НА РАЗНООБРАЗИЕ АНСАМБЛЯ

Мангалова Е. С., Мангалова М. С.
научный руководитель канд. техн. наук Шестернева О.В.
Сибирский государственный аэрокосмический университет им. М.Ф. Решетнева

В последние годы коллектив моделей (ансамбль) – один из распространенных инструментов решения задач интеллектуального анализа данных: классификации, идентификации, прогнозирования временных рядов.

Ансамблем $H(\bar{x})$ моделей $h_i(\bar{x})$ $(i=1,2,...N)$ называется композиция алгоритмических операторов $h_i : R^d \to R$ и корректирующей операции $F : R^N \to R$, в которой множеству оценок $h_1(\bar{x}), h_2(\bar{x}), ..., h_N(\bar{x})$ ставится в соответствие итоговая оценка $H(\bar{x})$ [1]:

$$H(\bar{x}) = F\left(h_1(\bar{x}), h_2(\bar{x}), ..., h_N(\bar{x})\right).$$

Фундаментальной задачей при построении ансамблей является генерация разнообразия ансамбля (или различия индивидуальных моделей) [2]. Очевидно, что агрегация схожих моделей в ансамбле не может привести к существенному повышению качества идентификации.

Проблема генерации разнообразия заключается в том, что индивидуальные модели обучаются для решения одной задачи, по одной обучающей выборке и вследствие этого обычно сильно коррелированны [3].

Предположим, что для аппроксимации функции $f : R^d \to R$ используются N индивидуальных моделей $h_1(\bar{x}), h_2(\bar{x}), ..., h_N(\bar{x})$, объединенных в ансамбль:

$$H(\bar{x}) = \frac{1}{N}\sum_{i=1}^{N} h_i(\bar{x}),$$

а для оценки качества ансамбля используется квадратичная ошибка:

$$E(H) = \frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N}\left(f(\bar{x}_j) - H(\bar{x}_j)\right)^2.$$

Одной из наиболее распространенных характеристик точности моделей в ансамбле является средняя ошибка индивидуальных моделей в ансамбле:

$$E_m(H) = \sum_{i=1}^{N}\frac{1}{nN}\sum_{j=1}^{n}\left(h_i(\bar{x}_j) - f(\bar{x}_j)\right)^2.$$

Разница между квадратичной ошибкой ансамбля и средней ошибкой индивидуальных моделей показывает то, насколько разнообразны входящие в ансамбль модели:

$$E(H) = E_m(H) - A(H), \qquad (1)$$

$$A(H) = \frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N}\left(h_i(\bar{x}_j) - H(\bar{x}_j)\right)^2$$

Декомпозиция (1) была предложена Крогом и Веделсби в работе [4]. Эта декомпозиция показывает влияние точности индивидуальных моделей и их разнообразия на ошибку коллективной модели. В силу того, что компонента $A(H)$ неотрицательна, ошибка ансамбля $E(H)$ не может быть больше, чем средняя ошибка

индивидуальных моделей $E_m(H)$. Следовательно, наилучший ансамбль состоит из более точных и одновременно более разнообразных индивидуальных моделей.

Решение задачи максимизации разнообразия ансамбля при его построении является одним из важнейших вопросов теории и практики коллективных методов анализа данных.

Из обучающей выборки могут быть сформированы различные обучающие подмножества. Чем меньше размерность этих подмножеств, тем меньше мощности попарных пересечений этих подмножеств, а следовательно, тем более разнообразными будут обученные на них индивидуальные модели. Одновременно с этим, уменьшение размерности обучающих подмножеств неизбежно будет приводить к уменьшению точности индивидуальных моделей.

Покажем влияние размерности обучающих подмножеств и способа их формирования на разнообразие ансамбля.

В качестве тестовой задачи были взяты данных по жилищному фонду Бостона [5]. Отклик: медианное значение стоимости различных домов в тысячах долларов. Предикторы: 13 характеристик недвижимости и ее местоположения. Объем обучающей выборки – 450, объем тестовой выборки – 56.

Генерация обучающих подмножеств случайным образом. Из обучающей выборки формируется N подмножеств размерностью M: вероятность попадания каждого наблюдения в i-е обучающее подмножество – M / n.

На рис. 1 приведена зависимость ошибки ансамбля от меры разнообразия при изменении размерности обучающих множеств M, на рис. 2 – зависимость средней меры разнообразия $A(H)$ от M. С ростом M увеличиваются мощности пересечения обучающих подмножеств, а модели, построенные на этих подмножествах, становятся более похожими.

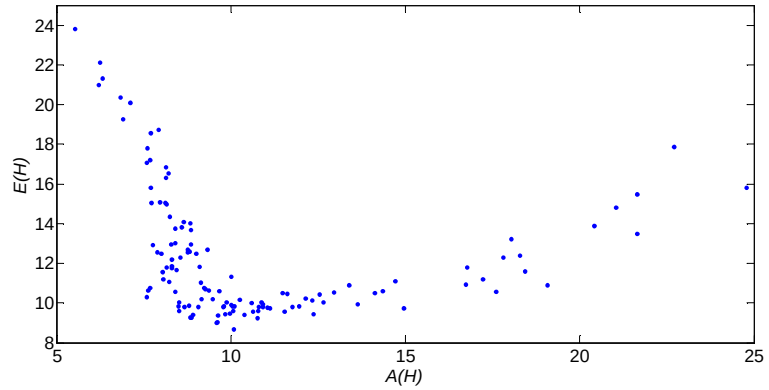

Рис. 1. Зависимость ошибки ансамбля от его разнообразия. Обучающие подмножества генерируются случайным образом

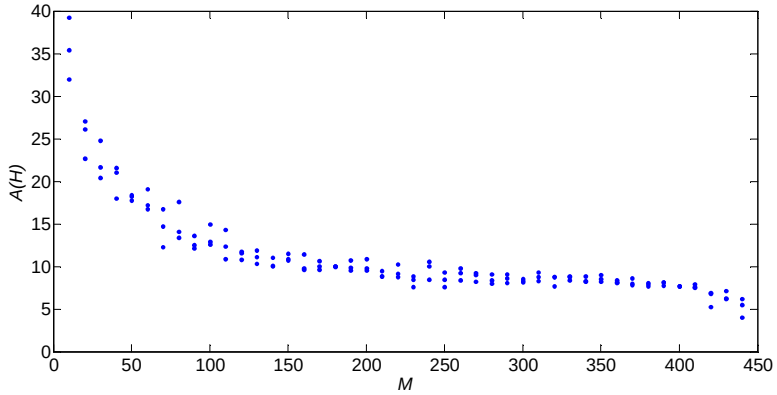

Рис. 2. Зависимость разнообразия ансамбля от размерности обучающего подмножества

Наилучшие значения параметра M для данного способа формирования обучающих подмножеств принадлежат интервалу [100, 200].

Генерация обучающих подмножеств на основании ошибок текущего ансамбля. В отличие от предыдущего метода процесс формирования обучающих подмножеств носит итеративный характер. Первое обучающее подмножество размерностью M формируется случайным образом. Последующие подмножества $(i = 2,3,...,N)$ составляются из M наблюдений, для которых ошибки текущего ансамбля $H_i(\bar{x})$ наибольшие.

На рис. 3 приведена зависимость ошибки ансамбля $E(H)$ от меры разнообразия $A(H)$. Наилучшие значения размерности обучающих подмножеств M принадлежат интервалу [100, 250].

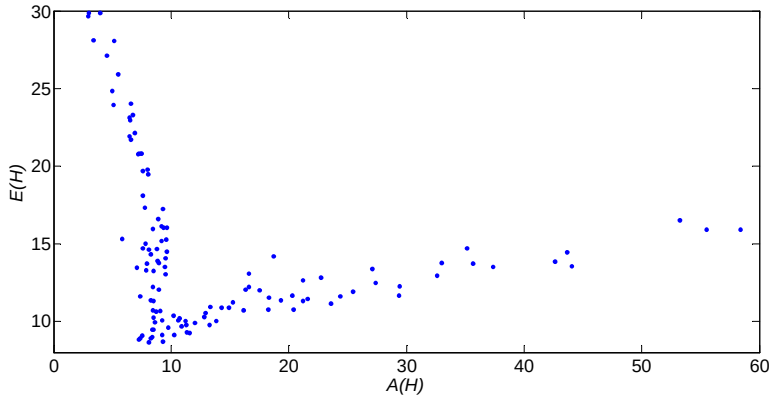

Рис. 3. Зависимость ошибки ансамбля от его разнообразия. Обучающие подмножества генерируются на основании ошибок текущего ансамбля

Использование сведений об ошибках ансамбля на каждой итерации позволяет скорректировать точность и разнообразие ансамбля на следующей итерации. Тем не менее, выбор размерности обучающих подмножеств оказывает большее влияние на точность ансамбля, чем способ их формирования.

Список литературы

1. Журавлёв Ю. И., Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. 1978. Т.33. С. 5-68.
2. Kuncheva L. I. Combining Pattern Classifiers: Methods and algorithms. John Wiley & Sons, Hoboken, NJ, 2004.
3. Zhou Z.-H. Ensemble Methods: Foundations and algorithms. Chapman & Hall/Crc Machine Learning & Pattern Recognition. 2012. 236 p.
4. Krogh A., Vedelsby J. Neural network ensembles, cross validation and active learning // Advanced in Neural Information Processing System 7. Cambridge: MIT press, 1995. P. 231-238.
5. Harrison D., Rubinfeld D.C. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5, 81-102.